WDS'10 Proceedings of Contributed Papers, Part I, 88–93, 2010. ISBN 978-80-7378-139-2 © MATFYZPRESS

English-Czech Machine Translation Using TectoMT

M. Popel

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. English to Czech machine translation as it is implemented in the TectoMT system consists of three phases: analysis, transfer and synthesis. The system uses tectogrammatical (deep-syntactic dependency) trees as the transfer medium. Each phase is divided into so-called blocks, which are processing units that solve linguistically interpretable tasks (e.g., statistical part-of-speech tagging or rule-based placement of clitics).

This paper shortly introduces linguistic layers of language description which are used for the translation and describes basic concepts of the TectoMT framework. The translation results are evaluated using both automatic metric BLEU and human judgments from the WMT 2010 evaluation.

1. Introduction

Machine translation (MT) is gaining more and more importance in the contemporary world. There are many approaches to MT, which are by tradition classified into two paradigms: rule-bases and statistical.

Classical rule-based MT systems make use of linguistic knowledge (grammars, dictionaries, rules written by human experts), but they use no information learned automatically from corpora. The translation usually comprises three phases: analysis, transfer and synthesis. MT systems can be further classified according to the level of language abstraction used for the transfer – some systems perform shallow analysis, some perform deep (or rich) analysis.[1] The advantage of deeper analysis is that the transfer should be easier and when building a system for translation between more than two languages, the analysis and synthesis can be shared for all language pairs with the given source or target language respectively.

Classical statistical MT systems make use of large-scale human-translated parallel corpora and monolingual corpora, but almost no linguistic knowledge. This has the advantage that the same system architecture can be used for any pair of languages for which there are enough training data available.

In recent years, there is a tendency to exploit linguistic knowledge to a greater extend to improve the performance of statistical MT systems. On the other hand, rule-based or syntax-based MT systems incorporate more statistical methods (rules can be automatically learned from parallel corpora, stochastic taggers and parsers are being used etc.). This results in a convergence of both the paradigms.

This paper describes improvements of English-Czech translation system called TectoMT. TectoMT[2] is one of the promising MT systems that combine statistical techniques and linguistic knowledge. It aims at transfer on the so-called tectogrammatical layer, which is a layer of deep syntactic dependency trees.

The rest of the paper is organized as follows: Section 2 introduces the system of linguistic layers, Section 3 describes basic concepts of the TectoMT framework, and Section 4 briefly overviews the translation process. Finally, we conclude with evaluation in Section 5.

[1] Shallow syntax structure of a sentence can be represented either by constituency trees [e.g. Wang et al., 2007] or dependency trees [e.g. Quirk et al., 2005]. For a deep syntax structure it is more common to use dependency trees such as tectogrammatical trees from the Functional Generative Description theory [Sgall, 1967], normalized trees in the ETAP-3 system [Boguslavsky et al., 2004], or logical form structures [Menezes and Richardson, 2001].

[2] `http://ufal.mff.cuni.cz/tectomt/`

2. Layers of Language Description

TectoMT profits from the stratificational approach to the language, namely it defines four layers of language description (listed in the order of increasing level of abstraction): raw text (word layer, w-layer), morphological layer (m-layer), shallow-syntax layer (analytical layer, a-layer), and deep-syntax layer (layer of linguistic meaning, tectogrammatical layer, t-layer).

The strategy is adopted from the Functional Generative Description theory [Sgall, 1967], which has been further elaborated and implemented in the Prague Dependency Treebank (PDT) [Hajič et al., 2006]. We give here only a very brief summary of the key points.

- **morphological layer (m-layer)**
  Each sentence is tokenized and each token is annotated with a lemma and morphological tag.

- **analytical layer (a-layer)**
  Each sentence is represented as a shallow-syntax dependency tree (a-tree). There is one-to-one correspondence between m-layer tokens and a-layer nodes (a-nodes). Each a-node is annotated with the so-called *analytical function*, which represents the type of dependency relation to its parent (i.e. its governing node).

- **tectogrammatical layer (t-layer)**
  Each sentence is represented as a deep-syntax dependency tree (t-tree). Autosemantic (meaningful) words are represented as t-layer nodes (t-nodes). Information conveyed by functional words (such as auxiliary verbs, prepositions and subordinating conjunctions) is represented by attributes of t-nodes. Most important attributes of t-nodes are: tectogrammatical lemma, functor (which represents the semantic value of syntactic dependency relation) and a set of grammatemes (e.g. tense, number, verb modality, deontic modality, negation). With regards to the needs of MT, some special attributes were added, most notably *formemes* (see Žabokrtský et al. [2008]).

3. TectoMT

TectoMT is a highly modular software framework for Natural Language Processing (NLP), implemented in Perl programming language under Linux. Although machine translation is the primary application of TectoMT, it is designed in a general way and it was already used for other tasks, e.g.: aligning tectogrammatical structures of parallel Czech and English sentences [Mareček et al., 2008], building a large, automatically annotated parallel English-Czech treebank CzEng 0.9 [Bojar and Žabokrtský, 2009], evaluating metrics for measuring translation quality [Kos and Bojar, 2009], or tagging the Czech data set for the CoNLL Shared Task [Hajič et al., 2009].

3.1. Blocks and Scenarios

Following the fundamental assumption that every non-trivial NLP task can be decomposed into a sequence of subsequent steps, these steps are implemented as reusable components called *blocks*. Each block has a well defined (and documented) input and output specification and also a linguistically interpretable functionality in most cases. This facilitates rapid development of new applications by simply listing the names of existing blocks to be applied to the data. Moreover, blocks in this sequence (which is called *scenario*) can be easily substituted with an alternative solution (other blocks), which attempts at solving the same subtask using a different approach or method.[3]

[3]Scenarios can be adjusted also by specifying parameters for individual blocks. Using parameters, we can define, for instance, which model should be used for parsing.

For example, the task of morphological and shallow-syntax analysis (and disambiguation) for English text consists of five steps: sentence segmentation, tokenization, part-of-speech tagging, lemmatization and parsing. In TectoMT we can arrange various scenarios to solve this task, for example:

Scenario A	Scenario B
`Sentence_segmentation_simple`	`Each_line_as_sentence`
`Penn_style_tokenization`	`Tokenize_and_tag`
`TagMxPost`	
`Lemmatize_mtree`	`Lemmatize_mtree`
`McD_parser`	`Malt_parser`

In the scenario A, tokenization and tagging is done separately in two blocks (`Penn_style_tokenization` and `TagMxPost`, respectively), whereas in the scenario B, the same two steps are done in one block at once (`Tokenize_and_tag`). Also different parsers are used.[4]

TectoMT currently includes over 400 blocks – approximately 140 blocks are specific for English, 120 for Czech, 60 for English-to-Czech transfer, 30 for other languages and 50 blocks are language-independent. Some of them contain only few lines of code, some solve complex linguistic phenomena. In order to prevent code duplications, many tools and routines are implemented as separate modules, which can be used in more blocks.

3.2. Documents, Bundles and Trees

Every document is saved in one file and consists of a sequence of sentences. Each sentence is represented by a structure called *bundle*, which stands for 'a bundle of trees'. Each tree can be classified according to:

- layer of language description (M=m-layer, A=a-layer, T=t-layer),
- language (e.g. Arabic, Czech, English, German[5]),
- indication whether the sentence was created by analysis (S=source) or by transfer/synthesis (T=target).

In other words, each bundle contains trees that represent the same sentence in different languages, layers and source/target direction (hence sentences in multilingual documents are implicitly aligned, see Figure 1). TectoMT trees are denoted by the three coordinates, e.g. analytical layer representation of an English sentence acquired by analysis is denoted as `SEnglishA`, tectogrammatical layer representation of a sentence translated to Czech is denoted as `TCzechT`. The convention is extremely useful for a machine translation that follows the analysis-transfer-synthesis scheme as illustrated in Figure 2 using Vauquois diagram. Nevertheless, also the blocks for other NLP tasks can be classified according to the languages and layers on which they operate.

4. Translation scenario outline

We briefly describe the whole process of English-Czech translation in TectoMT.

[4] `Penn_style_tokenization` is a rule-based block for tokenization according to Penn Treebank guidelines (`http://www.cis.upenn.edu/~treebank/tokenization.html`). `TagMxPost` uses Adwait Ratnaparkhi's tagger [Ratnaparkhi, 1996]. `Tokenize_and_tag` uses Aaron Coburns `Lingua::EN::Tagger` CPAN module. `Lemmatize_mtree` is a block for English lemmatization handling verbs, noun plurals, comparatives, superlatives and negative prefixes. It uses a set of rules (about one hundred regular expressions inspired by `morpha` [Minnen et al., 2000]) and a list of words with irregular lemmatization. `McD_parser` uses MST parser 0.4.3b [McDonald et al., 2005], `Malt_parser` uses Malt parser 1.3.1 [Nivre et al., 2007].

[5] In the near future, TectoMT will migrate to using ISO 639 language codes (e.g. ar, cs, en, de) instead of full names.

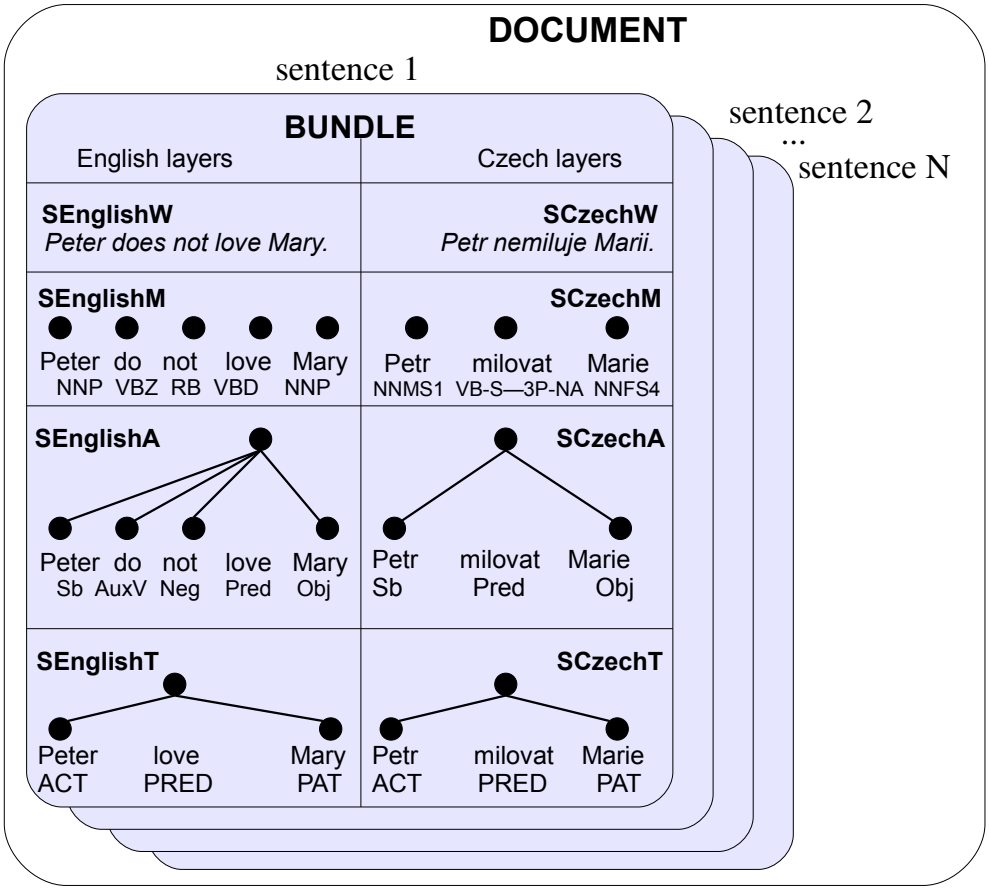

Figure 1. English-Czech parallel text annotated on three layers of language description (not counting the original text stored in the w-layer) is saved in a TectoMT document, each sentence in one bundle. We show only a simplified representation of the trees in the first bundle.

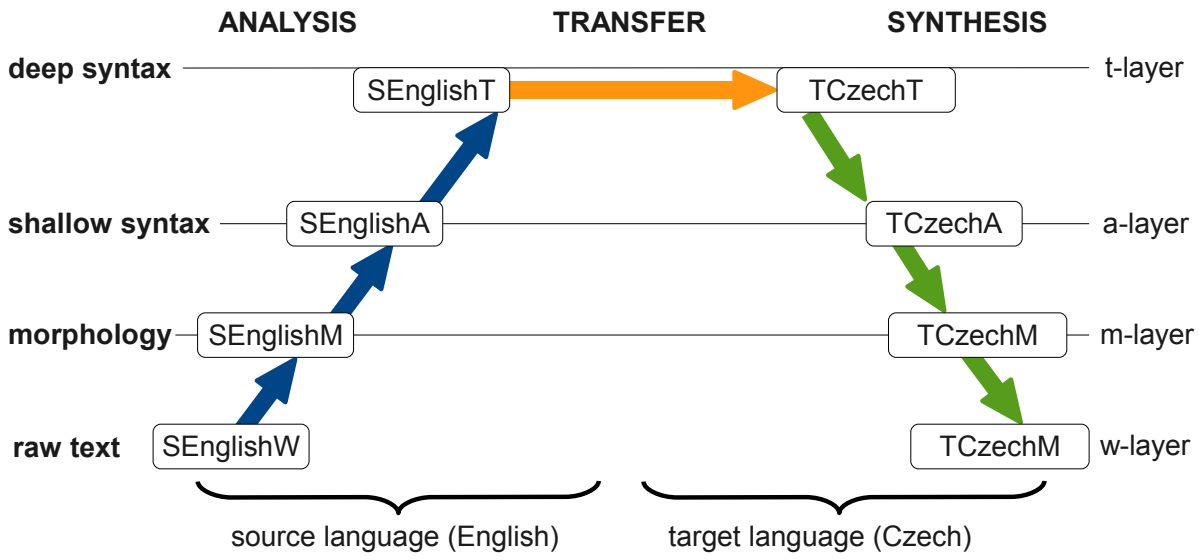

Figure 2. Vauquois diagram for translation with transfer on the tectogrammatical layer

4.1. Analysis from a raw text to English t-layer

In order to build the source m-layer, the English source text is segmented to sentences, tokenized, tagged using a part-of-speech tagger and lemmatized (see Scenario A in Section 3.1). The a-layer is build using the Maximum Spanning Tree Parser [McDonald et al., 2005] and a rule-based block for assignment of analytical functions. To convert a-trees to t-trees, we remove functional words (such as prepositions, subordinating conjunctions, articles etc.) and we encode the information conveyed in these words (and in word order and some morphological categories) into t-layer attributes (grammatemes, functor, semantic PoS, formeme and others).

4.2. Transfer from English t-layer to Czech t-layer

English tectogrammatical trees are translated to Czech tectogrammatical trees. Probabilistic dictionaries provide n-best lists of lemmas and formemes. The combination that is optimal for the whole tree is selected using the HMTM (Hidden Tree Markov Models) transfer block (see Žabokrtský and Popel [2009] for details). Additional rule-based blocks are used to translate other t-layer attributes.

4.3. Synthesis from Czech t-layer to a raw text

In this step Czech analytical trees are created from the tectogrammatical ones (auxiliary nodes are added), but the process of synthesis continuously goes on (morphological categories are filled, word forms are generated), so in the last block, the sentence is generated by simply flattening the tree and concatenating the word forms.

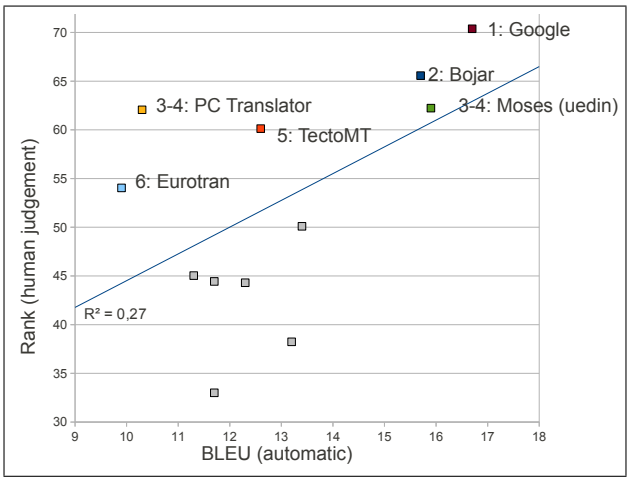

Figure 3. Correlation of human (ranks in %) and automatic (BLEU in %) evaluation of WMT 2010. Only the six best-ranked MT systems are shown with names.

5. Conclusion

Figure 3 shows results from the Fifth Workshop on Statistical Machine Translation – WMT 2010.[6] The participating systems were scored according to the human judgements (ranks) and according to automatic metric BLEU. Although the TectoMT system was ranked lower than state-of-the-art phrase-based systems (such as Google or Moses) and a commercial rule-based system PC Translator, it has the biggest BLEU improvement (from 7.3 to 12.6) compared with the last year (WMT 2009) among all the systems that participated in both the years. We hope for a similar progress in future.

[6] http://www.statmt.org/wmt10/

Acknowledgments. This work was supported by the grant GAUK 116310. I thank an anonymous reviewer for helpful comments.

References

Boguslavsky, I., Iomdin, L., and Sizov, V., Multilinguality in ETAP-3: Reuse of Lexical Resources, in *COLING 2004 Multilingual Linguistic Resources*, edited by G. Sérasset, pp. 1–8, COLING, Geneva, Switzerland, 2004.

Bojar, O. and Žabokrtský, Z., Building a Large Czech-English Automatic Parallel Treebank, *Prague Bulletin of Mathematical Linguistics*, *92*, 2009.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., and Mikulová, M., Prague Dependency Treebank 2.0, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 2006.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., and Zhang, Y., The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages, in *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*, Boulder, Colorado, USA, 2009.

Kos, K. and Bojar, O., Evaluation of Machine Translation Metrics for Czech as the Target Language, *Prague Bulletin of Mathematical Linguistics*, *92*, 2009.

Mareček, D., Žabokrtský, Z., and Novák, V., Automatic Alignment of Czech and English Deep Syntactic Dependency Trees, in *Proceedings of the Twelfth EAMT Conference*, edited by J. Hutchins and W. Hahn, pp. 102–111, HITEC e.V., Hamburg, 2008.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J., Non-Projective Dependency Parsing using Spanning Tree Algorithms, in *Proceedings of Human Langauge Technology Conference and Conference on Empirical Methods in Natural Language Processing (HTL/EMNLP)*, pp. 523–530, Vancouver, BC, Canada, 2005.

Menezes, A. and Richardson, S. D., A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora, in *Proceedings of the workshop on Data-driven methods in machine translation*, vol. 14, pp. 1–8, 2001.

Minnen, G., Carroll, J., and Pearce, D., Robust Applied Morphological Generation, in *Proceedings of the 1st International Natural Language Generation Conference*, pp. 201–208, Israel, 2000.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kbler, S., Marinov, S., and Marsi, E., MaltParser: A language-independent system for data-driven dependency parsing, *Natural Language Engineering*, *13*, 95–135, 2007.

Quirk, C., Menezes, A., and Cherry, C., Dependency treelet translation: Syntactically informed phrasal smt, in *ACL*, pp. 271–279, Ann Arbor, Michigan, 2005.

Ratnaparkhi, A., A maximum entropy part-of-speech tagger, in *Proceedings of the conference on empirical methods in natural language processing*, vol. 1996, pp. 133–142, 1996.

Sgall, P., *Generativní popis jazyka a česká deklinace*, Academia, Prague, Czech Republic, 1967.

Wang, W., Knight, K., and Marcu, D., Binarizing syntax trees to improve syntax-based machine translation accuracy, in *Proceedings of EMNLP and CoNLL 2007*, pp. 746–754, 2007.

Žabokrtský, Z. and Popel, M., Hidden Markov Tree Model in Dependency-based Machine Translation, in *Proceedings of ACL*, pp. 145–148, Suntec, Singapore, 2009.

Žabokrtský, Z., Ptáček, J., and Pajas, P., TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer, in *Proceedings of the 3rd Workshop on Statistical Machine Translation, ACL*, 2008.